PCT/DE 00/01791

BUNDESREPUBLIK DEUTSCHLAND

PRIORITY DOCUMENT
SUBMITTED OR TRANSMITTED IN
COMPLIANCE WITH
RULE 17.1(a) OR (b)

DE00/01791

Prioritätsbescheinigung über die Einreichung einer Patentanmeldung

ETW

Aktenzeichen:	199 33 984.8
Anmeldetag:	20. Juli 1999
Anmelder/Inhaber:	Siemens Aktiengesellschaft, München/DE
Bezeichnung:	Verfahren zur Bildung und/oder Aktualisierung von Wörterbüchern zum automatischen Adresslesen
IPC:	G 06 F, G 06 K

Die angehefteten Stücke sind eine richtige und genaue Wiedergabe der ursprünglichen Unterlagen dieser Patentanmeldung.

München, den 17. August 2000
Deutsches Patent- und Markenamt
Der Präsident
Im Auftrag

Jooss

A 9161
06/00
EDV-L

Beschreibung

Verfahren zur Bildung und/oder Aktualisierung von Wörterbüchern zum automatischen Adreßlesen

Die Erfindung betrifft ein Verfahren zur Bildung und/oder Aktualisierung von Wörterbüchern zum Adreßlesen.

Adreßlesesysteme benötigen Informationen über Inhalt und Syntax von Adressen, um die erforderlichen Informationen wie Stadt, Postleitzahl, Vorname und Nachname, etc. extrahieren zu können. Der zulässige Inhalt einzelner Adreßelemente wird mit einem Wörterbuch (Liste von zulässigen Zeichenketten) beschrieben, das nach dem Stand der Technik aus vorliegenden Informationsquellen aufgebaut wird, wie z.B. aus einem postalischen Wörterbuch oder aus einer Mitarbeiterliste einer Firma. Die Anwendungsdomäne ändert sich jedoch mit der Zeit, so daß das zu Beginn erstellte Wörterbuch nicht mehr alle vorkommenden Inhalte vollständig umfaßt. Vor allem bei der Anwendung eines Lesesystems zur innerbetrieblichen Postverteilung ist die Änderung des Wortvorrats beträchtlich: Mitarbeiter verlassen die Firma, neue Mitarbeiter kommen hinzu, Mitarbeiter wechseln die Abteilung oder Nachnamen ändern sich aufgrund von Heirat, etc. So fehlen im Wörterbuch Einträge und es gibt Einträge, die nicht mehr gültig sind. Je deutlicher der aktuell verwendete Wortvorrat vom Lexikon abweicht, desto mehr sinkt die Erkennungsleistung des Lesesystems.

Diese Änderungen mußten bisher in bestimmten Zeitabständen manuell in die Wörterbücher übertragen werden, so daß die geschilderten Nachteile auftraten.

Aufgabe der Erfindung ist es, ein Wörterbuch zum Adreßlesen automatisch zu bilden und/oder automatisch zu aktualisieren.

Erfindungsgemäß wird die Aufgabe durch die Merkmale des Anspruches 1 gelöst. Dabei wird von dem Gedanken ausgegangen, die Ergebnisse der aktuellen Leseprozesse zwischenzuspeichern,

auszuwerten und zum automatischen Aufbau oder zur Aktualisierung eines Wörterbuches zu nutzen. Beim Zwischenspeichern erfolgt eine Kennzeichnung, ob die jeweilige Adresse erfolgreich gelesen wurde oder ob sie zurückgewiesen wurde. Soll ein Wörterbuch neu erstellt werden oder sollen in das vorhandene Wörterbuch neue Adressaten aufgenommen werden, so werden die zurückgewiesenen Leseergebnisse herangezogen.

Die Wörterbücher können einzelne Wörter, z.B. Nachnamen und/oder zusammenhängende Wortgruppen mit n Wörtern, z.B. Vor- und Nachnamen oder Vor- und Nachnamen und Straßennamen enthalten, wobei die Wörter sowohl direkt nebeneinander (Abstand m=0) liegen als auch durch m Wörter beabstandet sein können.

Durch die Bildung von Klassen von Wörtern oder Wörtergruppen, die ein festgelegtes Mindestähnlichkeitsmaß zueinander besitzen, und die Aufnahme mindestens des Repräsentanten in das oder die Wörterbücher der zugeordneten Adreßbereiche, ist ein automatischer Aufbau eines Wörterbuches bzw. eine automatische Aktualisierung des Wörterbuches infolge neuer Adressaten oder von Änderungen bei den Adressaten möglich.

Vorteilhafte Ausgestaltungen der Erfindung sind in den Unteransprüchen beschrieben.

Zur Klassenbildung ist es vorteilhaft, eine Liste aller Wörter/Wortgruppen der zurückgewiesenen Leseergebnisse zu erstellen, die nach der Häufigkeit der Wörter/Wortgruppen sortiert ist. Dann wird, beginnend mit dem häufigsten Wort/Wortgruppe, das Ähnlichkeitsmaß mit allen übrigen Wörtern/Wortgruppen bestimmt und in eine Ähnlichkeitsliste eingetragen. Alle Wörter/Wortgruppen in der Ähnlichkeitsliste mit einem Ähnlichkeitsmaß über einer festgelegten Schwelle werden anschließend dem aktuellen Wort/Wortgruppe als Klasse zugeordnet. Danach werden die Wörter/Wortgruppen der gebildeten Klasse aus der Häufigkeitsliste entfernt.

Die Repräsentanten der jeweiligen Klasse von Wörtern oder Wortgruppen der zwischengespeicherten und zurückgewiesenen Leseergebnisse können durch die kürzesten oder häufigsten Wörter oder Wortgruppen gebildet werden.

Zur Erkennung von Adressen im Wörterbuch, die geändert oder entfernt werden müssen, ist es vorteilhaft, die eindeutig gelesenen Adressen statistisch auszuwerten. Tritt eine plötzliche Änderung der Häufigkeit der Wörter und/oder Wortgruppen über eine bestimmte Schwelle hinaus auf und dauert sie eine festgelegte Zeit an, so werden diese Wörter/Wortgruppen aus dem Wörterbuch entfernt.

Um zu vermeiden, daß irrelevante Wörter der Leseergebnisse in das Wörterbuch aufgenommen werden, können diese durch Vergleich mit in einer speziellen Datei für irrelevante Wörter gespeicherten Wörtern ermittelt werden.

Vorteilhaft in diesem Zusammenhang ist es auch, kurze Wörter ohne Abkürzungspunkt mit weniger als p Buchstaben als irrelevant nicht ins Wörterbuch aufzunehmen. Um die Adreßinterpretation mit Hilfe der Wörterbücher möglichst detailliert durchzuführen, ist es vorteilhaft, neben den Repräsentanten auch die Wörter und/oder Wortgruppen der dazugehörenden Klassen mit den Ähnlichkeitsmaßen und Häufigkeiten aufzunehmen.

In einer weiteren vorteilhaften Ausgestaltung können zusammengehörende Wortgruppen mit n Wörtern, die untereinander einen Abstand von m Wörtern haben, ermittelt werden, indem ausgehend vom jeweiligen, für das Wörterbuch ermittelten Einzelwort die Adressen mit Fenstern der Breite von n+m Wörtern durchsucht werden. Nachdem die weiteren n-1 Einzelwörter mit den Abständen von m Wörtern untereinander ermittelt wurden, erfolgt die Aufnahme dieser Wortgruppe mit ihren Häufigkeiten in das entsprechende Wörterbuch.

Vorteilhaft ist es auch, das Ähnlichkeitsmaß mit dem Levenshtein-Verfahren (siehe „A Method for the Correction of Garbled Words, Based on the Levenshtein Metric", K. Okuda, E. Tanaka, T. Kasai, IEEE Transactions on Computers, Vol. c-25, No. 2, February 1976) zu ermitteln.

Es kann auch vorteilhaft sein, die ermittelten Wörterbuchaktualisierungen an einem Videocodierplatz zu kategorisieren und bestätigen zu lassen oder die Neueintragungen ins Wörterbuch zusätzlich vor ihrer Übernahme in die entsprechende Kategorie mit den Inhalten einer Datei zu vergleichen, in der charakteristische allgemeingültige Namen oder wenigstens Zeichenstrings bezogen auf die jeweilige Kategorie (Vorname, Nachname, Abteilung) gespeichert sind.

Anschließend wird die Erfindung in einem Ausführungsbeispiel anhand der Zeichnung näher erläutert. Ziel hierbei ist, bisher unbekannte Nachnamen (n=1) oder Paare unbekannter Vor- und Nachnamen (n=2) oder Nach- und/oder Vor- und Nachnamen und Abteilungsnamen von Mitarbeitern einer Firma und/oder entsprechende nicht mehr gültige Namen bzw. Namenskombinationen zu ermitteln und Wörterbuchänderungen durchzuführen.

Dabei zeigen

FIG 1 eine Ablaufstruktur eines Monitorprozesses zur Überwachung und Steuerung der Aktualisierung des Wörterbuches

FIG 2 eine Ablaufstruktur zur Ermittlung und Kennzeichnung irrelevanter Wörter

FIG 3 eine Ablaufstruktur zur Ermittlung bisher unbekannter Einzelwörter (n=1) (Nachnamen)

FIG 4 eine Ablaufstruktur zur Ermittlung bisher unbekannter Wortgruppen, ausgehend von den Einzelwörtern

FIG 5 eine Ablaufstruktur zur Aktualisierung der Wörterbücher unter Berücksichtigung der Wortkategorien

Die Wortvorschläge werden aus den Erkennungsergebnissen automatisch generiert, die das Lesesystem im täglichen Betrieb für jedes Sendungsbild berechnet. Die Erkennungsergebnisse für jedes Sendungsbild umfassen unterschiedliche geometrische Objekte (Layoutobjekte), wie Textblöcke, Zeilen, Wörter und Zeichen, und deren Relationen untereinander, also, welche Zeilen zu welchem Textblock gehören, welche Wörter in welchen Zeilen liegen, etc. Für jedes Einzelzeichenbild erzeugt das Lesesystem eine Liste von möglichen Zeichenbedeutungen. Darüberhinaus berechnet das Lesesystem für jedes Layoutobjekt seine Lage im Sendungsbild und dessen geometrischen Ausmaße.

Zum Aktualisieren oder auch Lernen von Wörterbucheinträgen wird die Menge der bearbeiteten Sendungen in zwei Teilmengen getrennt, in die Menge der vom Lesesystem automatisch gelesenen (aber nicht notwendigerweise korrekt gelesenen) und die Menge der zurückgewiesenen Sendungen. Die Menge der automatisch gelesenen Sendungen dient zum Ermitteln von Wörterbucheinträgen, die nicht mehr gültig sind; aus der Menge der zurückgewiesenen Sendungen werden neue Wörterbucheinträge abgeleitet.

Das beispielhafte System besteht aus fünf Modulen: einen Monitorprozeß, einer Aufbereitung der Erkennungsergebnisse (Vorverarbeitung), zweier Wörterbuchgenerierungsverfahren und einem Vorschlagsadministrator.

Der Monitorprozeß gemäß FIG 1 überwacht und steuert das Wörterbuchlernen. Die Erkennungsergebnisse 21 für jedes Sendungsbild werden zusammen mit einer Kennung für „erfolgreich gelesen" oder „zurückgewiesen" vom Leser an den Monitor übergeben. Zusätzliche Informationen zur Sendungsart (Brief, Großbrief, Hauspostformular) und weitere Merkmale zu den einzelnen Objekten der Erkennungsergebnisse, wie ROI (Region of Interest), Zeilen- und Wort-Hypothesen, Zerlegungsalternativen und Schriftzeichen-Erkennungsergebnisse, können ebenfalls übergeben werden. Diese Erkennungsergebnisse werden im Monitor in einem Zwischenspeicher 22 gespeichert, bis eine genügend große Menge an Daten angefallen ist (z.B. nach 20.000 Sendungen oder nach einer Woche Betrieb).

Im einfachsten Fall wird lediglich die erste Alternative der Zeichenerkennungsergebnisse zusammen mit dem besten Segmentierpfad im Zwischenspeicher gespeichert. Beispielsweise könnte der Inhalt folgendermaßen aussehen:

<Erkennungsergebnisse>	<Kennung>
:...	
1017921 PMD 55	erkannt
MR. A1FRED C SCHMIDI EXCCU1LVE DIRCC1OR, 0PCRA1IONS DCVC1OPMENT MyComp, INC 1 MyStreet MyCity, 12345	
PO11Y O/BRIEN	zurückgewiesen, nicht im Wörterbuch
MANAGER, COMMUNITY AFFAIRS MyComp INC 1 MyStreet MyCity, 12345	
PO1LY OBRIEN	zurückgewiesen, nicht im Wörterbuch
MANAGER, COMMUNITY AFFAIRS MyComp, INC 1 MyStreet MyCity, 12345	
MS ME1INDA DUCKSWORTH MyComp, INC MAI1 CODE 63-33 1 MyStreet MyCity, 12345	erkannt

**********AUR0**MIXED AADC 460 zurückgewiesen, nicht im Wörterbuch

MIK0 SCHWARTZ
O AND T 26-OO
1 MyStreet
MyCity, 12345

.....

Liegen genügend Ergebnisse vor, werden die zurückgewiesenen Erkennungsergebnisse an eine Aufbereitungseinheit 30 übergeben und zu den beiden Teilprozessen zum Wörterbuchlernen für Einzelworte 50 und Wortgruppen 60 weitergeleitet. Im Falle einer erfolgreichen automatischen Erkennung werden die Ergebnisse an ein Statistikmodul übergeben 40. Wenn alle Sendungen verarbeitet worden sind, werden die Wort- und Wortgruppenlisten 41 des Statistikmoduls und der Wörterbuchlernprozesse 51, 61 gesammelt und mit einer geeigneten grafischen Oberfläche einer Bedienkraft zur Bestätigung vorgelegt.

In der Aufbereitungseinheit 30 werden irrelevante Wörter in den zurückgewiesenen Erkennungsergebnissen gekennzeichnet, die in der nachfolgenden Textanalyse nicht berücksichtigt werden (vgl. FIG 2). Diese Wörter werden als nicht relevant markiert aber nicht gelöscht, da die Wortnachbarschaft für den nachfolgenden Wörterbuchaufbau wichtig ist.

Im Verfahrensschritt Markieren irrelevanter Wörter 31, werden aus der Menge der Worthypothesen kurze Wörter markiert, beispielsweise diejenigen, die weniger als 4 Buchstaben lang sind und gleichzeitig keinen Abkürzungspunkt besitzen, und solche die zu weniger als 50% aus alphanumerischen Zeichen bestehen. Weiterhin werden solche Wörter markiert, die in einer speziellen Datei 32 enthalten sind, die für diese Anwendung häufige, aber irrelevante Wörter enthält. Bei der Anwendung der innerbetrieblichen Postverteilung können beispiels-

weise der Firmenname, Städtename, Straßenname, Postfachbezeichnung, etc., in diesem speziellen Lexikon enthalten sein. Die Ergebnisse der Aufbereitung werden in einen Zwischenspeicher 33 zurückgeschrieben.

Nach der Vorverarbeitung sehen die Ergebnisse folgendermaßen aus:

<title MR> <first-name ALFRED> <last-name SCHMID>
<role EXECUTIVE DIRECTOR OPERATIONS>

PO11Y O/BRIEN
MANAGER, COMMUNITY AFFAIRS
<irrelevant MyComp, INC>
<irrelevant 1 MyStreet>
<irrelevant MyCity> <irrelevant 12345>

PO1LY OBRIEN
MANAGER, COMMUNITY AFFAIRS
<irrelevant MyComp, INC>
<irrelevant 1 MyStreet>
<irrelevant MyCity> <irrelevant 12345>

<title MS> <first-name MELINDA> <last-name DUCKSWORTH>

<non-alpha **********AUR0**MIXED> AADC <short 46O>
MIK0 SCHWARTZ
<short O> <short AND> <short T> 26 OO
<irrelevant MyComp, INC>
<irrelevant 1 MyStreet>
<irrelevant MyCity> <irrelevant 12345>

.....

Aus den aufbereiteten zurückgewiesenen Erkennungsergebnissen wird gemäß FIG 3 im ersten Schritt 52 eine Häufigkeitsliste FL 53 aller darin vorkommender Wörter erstellt, nach abstei-

gender Häufigkeit sortiert und in einen Zwischenspeicher abgelegt. Für obiges Beispiel könnte die Häufigkeitsliste FL 53 folgendermaßen aussehen:

...	
AFFAIRS	37
MANAGER	37
COMMUNITY	37
OBRIEN	20
O/BRIEN	17
SCHWARTZ	15
MIK0	12
POLLY	10
PO11Y	8
PAULA	8
PO1LY	5
MIKO	3
.....	

Aus dieser Liste wird schrittweise ein Wörterbuch W1 relevanter Wörter 51 aufgebaut. Zu jedem Wort in der Häufigkeitsliste FL 53 wird der Abstand d zu allen Wörtern in dieser Häufigkeitsliste bestimmt. Ein Verfahren zur Messung des Abstandes zwischen zwei Zeichenketten ist das Levenshtein-Verfahren, das den minimalen Abstand zweier Zeichenketten berechnet, bezogen auf 3 Kostenarten, auf Kosten einer Ersetzung eines Zeichens, einer Einfüge- und einer Löschoperation. Zur Berechnung von d können neben der Zeichenkette weitere Merkmale der Erkennungsergebnisse verwendet werden, beispielsweise die Zeichenalternativen, die Segmentieralternativen, etc.

Das erste Wort in der Häufigkeitsliste FL 53 (das aktuell häufigste) wird in das Wörterbuch W1 51 übernommen und aus der Häufigkeitsliste FL 53 gelöscht 54. Alle Wörter aus der Häufigkeitsliste FL 53 mit einem Abstand kleiner einer fest-

gelegten Schwelle th_d werden dem aktuellen Wort im Wörterbuch W1 51 mit ihrer Häufigkeit zugeordnet 55, 56. Gleichzeitig werden diese Wörter in der Häufigkeitsliste FL 53 gelöscht. Die Iteration endet, wenn die Häufigkeitsliste FL 53 leer ist. Damit werden Wortklassen gebildet, die untereinander einen Abstand d nicht überschreiten, bzw. ein entsprechendes Ähnlichkeitsmaß nicht unterschreiten.

Wenn alle Wörter verarbeitet sind, besteht das Wörterbuch W1 51 aus einer Menge von Wortklassen. Das kürzeste Wort einer Wortklasse wird als Repräsentant der Gruppe bezeichnet. Jede Wortklasse enthält Wörter, die einander ähnlich sind, mit den dazugehörigen Häufigkeiten und Abständen zum Klassenrepräsentanten. Die Repräsentanten der Wortklassen im Wörterbuch W1 51, und damit auch die Wortklassen, werden nach absteigender Häufigkeit sortiert 57. Die Häufigkeit einer Wortklasse setzt sich aus der Häufigkeit des Repräsentanten und der Häufigkeiten der Elemente der Wortklasse zusammen. Wortklassen, deren Häufigkeit eine gewisse Schwelle unterschreiten, werden aus dem Wörterbuch W1 51 gelöscht. Aus obiger Liste wird folglich folgendes Wörterbuch W1 51 gebildet:

<Wortklasse>	<Häufigkeit>	<Abstand>
...		
AFFAIRS	37	
MANAGER	37	
COMMUNITY	37	
OBRIEN	37	
O/BRIEN	17	(d = 1)
POLLY	23	
PO11Y	8	(d = 2)
PO1LY	5	(d = 1)
SCHWARTZ	15	
MIKO	15	
MIKO	3	(d = 1)
PAULA	8	

...

Die Bildung von Repräsentanten kann je nach Anwendung mit weiterem Wissen unterstützt werden. So kann ein Wort entweder auf eine Zahl oder auf eine Alpha-Folge abgebildet werden, indem OCR-Ersetzungstabellen verwendet werden, die austauschbare Zeichenpaare definieren, wie 1 - L, 0 - O, 2 - Z, 6 - G, etc. Wenn darüberhinaus zu erlernenden Wortklassen Alternativenmengen bekannt sind - für Vornamen beispielsweise Spitznamen, wie Paula-Polly, Thomas-Tom, etc., kann auch diese Ersetzung vorgenommen werden. Beide Schritte können auf das Wörterbuch W1 51 angewendet werden, was zu einer weiteren Verschmelzung von Wortklassen führt.

Abschließend werden in den Erkennungsergebnissen alle Wörter, die im Wörterbuch W1 51 vorkommen, markiert und durch ihren Repräsentanten ergänzt. Diese Wörter werden im folgenden mit W1-Wörter bezeichnet.

An der Spitze vom Wörterbuch W1 51 stehen nun die häufigsten, bisher unbekannten Wortformen und die Wortklassen enthalten Schreibvarianten davon. So werden in der Anwendung der innerbetrieblichen Postverteilung bisher unbekannte Nach- und Vornamen und Teile von Abteilungsbezeichnungen im Wörterbuch W1 51 stehen. Darüberhinaus enthalten deren Wortklassen Schreibvarianten oder Varianten, die aufgrund der Eigenschaften des Lesesystems entstanden sind.

Ausgehend von den Repräsentanten der Wortklassen im Wörterbuch W1 51, die in den Erkennungsergebnissen als solche markiert sind, werden im nächsten Schritt nach FIG 4 Wortgruppen der Länge 2 bis n bestimmt, indem die Nachbarschaften von W1-Wörtern der Erkennungsergebnisse 62 untersucht werden. Für jedes W1-Wort wird dazu die rechte Nachbarschaft in einem Fenster der Breite k <= n durchsucht, ob darin weitere W1-Wörter sind. n-1 zunächst leere Wörterbücher werden in einem

Zwischenspeicher angelegt und Schritt für Schritt gefüllt. Ein n-Tupel wird dann in einen Wortgruppen-Zwischenspeicher aufgenommen 53, wenn n W1-Wörter gefunden worden sind und weniger als m weitere nicht W1-Wörter zwischen diesen n liegen. Wie beim Wörterbuch W1 51, wird auch hier die Auftretenshäufigkeit der einzelnen Wortgruppen der Länge n gespeichert.

Der Wahl der Werte von m und n hängt von der konkreten Anwendung ab. Für Werte n > 4 sind bei der Anwendung Adreßlesen keine signifikant häufigen Einträge mehr zu erwarten. m = 0 bedeutet, daß alle n W1-Wörter direkt aufeinanderfolgen. Gerade bei Paaren von Vornamen und Nachnamen kann jedoch ein zweiter Vorname hin und wieder die direkte Aufeinanderfolge unterbrechen, genauso wie Segmentierfehler der Lesemaschine vermeintliche Worthypothesen erzeugen können und damit eine direkte Aufeinanderfolge verhindern. Für die beschriebene Anwendung sind folglich m=1 und n=3 geeignete Werte.
In diesem Schritt werden folglich aus dem Wortgruppen-Zwischenspeicher n-1 Wörterbücher Wn 61 generiert, die häufige Wortsequenzen mit ihren Häufigkeiten für Paare, Triplets, etc. bis zu n-Tupel enthalten. In jedem Wörterbüch Wn 61 werden die Häufigkeiten der n-Tupel mit den Häufigkeiten der W1-Wörter der n-Tupel zu einer Maßzahl verrechnet. Jedes Wörterbuch Wn 61 wird nach absteigenden Maßzahlen sortiert, so daß wieder die signifikantesten Wortgruppen am Anfang eines jeden Wörterbuches Wn stehen 54.

Für obiges Beispiel sieht das Wörterbuch W2 folgendermaßen aus:

W2

COMMUNITY AFFAIRS	37
MANAGER COMMUNITY	37
POLLY OBRIEN	23
MIKO SCHWARTZ	15
PAULA OBRIEN	8

Das Wörterbuch W3 hat 3 Einträge, vorausgesetzt, daß der Name POLLY OBRIEN stets mit der Bezeichnung MANAGER COMMUNITY AFFAIRS kombiniert vorkommt und ein Zeilenumbruch in einem n-Tupel erlaubt ist.:

W3

MANAGER COMMUNITY AFFAIRS	37
POLLY OBRIEN MANAGER	23
OBRIEN MANAGER COMMUNITY	23

Wie beschrieben werden nun die Wortvorschläge der Wörterbücher Wn 61 (W2, W3, etc) entsprechend FIG 5 einem Operator zur Validierung vorgelegt. Durch Wissen über die zu erlernenden Worteinheiten 72 ist es an dieser Stelle möglich, Einträge in den Wörterbüchern W1, W2, .. Wn 51, 61 semantisch zu kategorisieren 71. So lassen sich in dieser Anwendung Einträge der semantischen Klasse <Name> zuordnen, indem in allgemeingültigen Vornamenslisten nachgeschlagen wird. Ähnliches gilt für die Semantikklasse <Abteilung>, die sich aus Schlüsselwörtern wie Department ableiten läßt.
Dieser Vorgang ist selbstverständlich auch automatisch ohne Operator durch Vergleich mit den Einträgen dieser Listen auszuführen.

Zu erfolgreich verteilten Sendungen sind die dazu erforderlichen Adreßelemente gefunden worden und sind als solche in den Erkennungsergebnissen gekennzeichnet. Wenn beispielsweise in der Anwendung der innerbetrieblichen Postverteilung Nachnamen und Vornamen erfolgreich gelesen worden sind, werden diese Ergebnisse in einer Statistik erfaßt; insbesondere wird die Häufigkeit der extrahierten Wörter, Paare, im allgemeinen von n-Tupeln, über definierte Zeitabschnitte td, z.B. für eine Woche, gespeichert, wobei die Sendungsart berücksichtigt werden kann. Als Ergebnis erhält man eine Verteilung der zu ex-

trahierenden Adreßelemente für eine Folge von Zeitabschnitten:

Zeitpunkt 1	
MELINDA DUCKSWORTH	123
ALFRED SCHMID	67
...	
Zeitpunkt 2	
MELINDA DUCKSWORTH	1
ALFRED SCHMID	85
...	
Zeitpunkt 3	
MELINDA DUCKSWORTH	2
ALFRED SCHMID	72
...	

Aus der so ermittelten Verteilung läßt sich ableiten, ob Wörterbucheinträge gelöscht werden sollen: Die Einträge werden in eine Liste zum Entfernen aus dem Wörterbuch eingefügt, wenn deren Häufigkeit sich von td_i zu td_{i+1} abrupt verringert und auf diesem Niveau in aufeinanderfolgenden Zeitabschnitte td_{i+k} bleibt (z.B. k = 4). So wird im obigen Beispiel die Person MELINDA DUCKSWORTH im Wörterbuch gelöscht. Dieser Ablauf kann zusätzlich auch über einen Bestätigungsvorgang geführt werden.

Patentansprüche

1. Verfahren zur Bildung und/oder Aktualisierung von Wörterbüchern zum automatischen Adreßlesen,
gekennzeichnet durch die Schritte:
- Zwischenspeicherung der vom OCR-Leser erzielten Leseergebnisse der Adressen einer vereinbarten Anzahl von Sendungsbildern oder innerhalb einer vereinbarten Zeitspanne gelesener Sendungsbilder, unterteilt in eindeutig gelesene Ergebnisse mit einer Übereinstimmung mit einem Wörterbucheintrag und in zurückgewiesene Leseergebnisse ohne Übereinstimmung mit einem Wörterbucheintrag,
- Bildung von Klassen von Wörtern mit dazugehörenden Repräsentanten oder zusammengehörenden Wortgruppen der zwischengespeicherten und zurückgewiesenen Leseergebnisse, bestehend jeweils aus n Adreßwörtern, n= 1,2,...a, mit den Wortabständen m, m= 0,1,...b, die bezogen auf jeweils einen bestimmten n- und m-Wert untereinander ein bestimmtes Ähnlichkeitsmaß nicht unterschreiten,
- Aufnahme mindestens der Repräsentanten derjenigen Klassen, deren Häufigkeit einen festgelegten Wert überschreiten, in das oder die Wörterbücher der zugeordneten Adreßbereiche.

2. Verfahren nach Anspruch 1, dadurch gekennzeichnet, daß
- zur Klassenbildung eine Häufigkeitsliste aller vorkommenden Wörter oder Wortgruppen der zurückgewiesenen Leseergebnisse, nach deren Häufigkeit sortiert, erstellt wird,
- zu jedem Wort oder jeder Wortgruppe, beginnend mit dem häufigsten Wort oder der häufigsten Wortgruppe, das Ähnlichkeitsmaß mit allen übrigen Wörtern oder Wortgruppen bestimmt und in eine Ähnlichkeitsliste eingetragen wird,
- alle Wörter oder Wortgruppen in der Ähnlichkeitsliste mit einem Ähnlichkeitsmaß über einer festgelegten Schwelle dem aktuellen Wort oder der aktuellen Wortgruppe als Klasse zugeordnet werden,

- anschließend die Wörter oder Wortgruppen der jeweils gebildeten Klasse aus der Häufigkeitsliste entfernt werden.

3. Verfahren nach Anspruch 1, d a d u r c h g e k e n n z e i c h n e t, daß
der Repräsentant der jeweiligen Klasse von Wörtern oder Wortgruppen der zwischengespeicherten und zurückgewiesenen Leseergebnisse durch das/die kürzeste oder häufigste Wort oder Wortgruppe gebildet wird.

4. Verfahren nach Anspruch 1, d a d u r c h g e k e n n z e i c h n e t, daß
die zeitliche Häufigkeit der Wörter oder Wortgruppen der eindeutig gelesenen Adressen statistisch dahingehend ausgewertet werden, daß bei deren plötzlicher und über einen festgelegten Zeitraum andauernder Verringerung über eine festgelegte Schwelle die jeweiligen eingetragenen Wörter oder Wortgruppen aus dem Wörterbuch entfernt werden.

5. Verfahren nach Anspruch 1, d a d u r c h g e k e n n z e i c h n e t, daß
irrelevante Wörter der Leseergebnisse durch Vergleich mit in einer speziellen Datei gespeicherten Wörtern ermittelt und nicht in das Wörterbuch aufgenommen werden.

6. Verfahren nach Anspruch 1, d a d u r c h g e k e n n z e i c h n e t, daß
kurze Wörter ohne Abkürzungspunkt mit weniger als p Buchstaben nicht in das Wörterbuch aufgenommen werden.

7. Verfahren nach Anspruch 1, d a d u r c h g e k e n n z e i c h n e t, daß
in das Wörterbuch neben den Repräsentanten auch die Wörter und/oder Wortgruppen der dazugehörenden Klassen mit den Ähnlichkeitsmaßen und Häufigkeiten eingetragen werden.

8. Verfahren nach einem der Ansprüche 1 und 2, d a d u r c h g e k e n n z e i c h n e t, daß
für Wörtergruppen mit n Wörtern, n>1, wobei die Wörter untereinander einen Abstand von m Wörtern, m>=0, haben, ausgehend vom jeweiligen, für das Wörterbuch ermittelten Einzelwort die Adressen mit Fenstern der Breite von n+m Wörtern durchsucht werden und beim Finden von weiteren n-1 für das Wörterbuch ermittelten Einzelwörtern in den festgelegten Abständen m untereinander diese gefundenen Wortgruppen mit deren Häufigkeiten in das entsprechende Wörterbuch übernommen werden.

9. Verfahren nach einem der Ansprüche 1,2,7,8, d a d u r c h g e k e n n z e i c h n e t, daß
das Ähnlichkeitsmaß zwischen den Wörtern mit dem Levenshtein-Verfahren ermittelt wird.

10. Verfahren nach einem der Ansprüche 1 bis 9, d a d u r c h g e k e n n z e i c h n e t, daß
die zu entfernenden Wörterbucheintragungen und die Neueintragungen ins Wörterbuch an einem Videocodierplatz angezeigt, kategorisiert und bestätigt werden.

11. Verfahren nach einem der Ansprüche 1 bis 9, d a d u r c h g e k e n n z e i c h n e t, daß
die ins Wörterbuch einzutragenden Wörter und/oder Wortgruppen vor deren Eintragung mit den Inhalten einer Datei verglichen werden, in der für die jeweilige Wörterbuchkategorie charakteristische, allgemeingültige Namen oder wenigstens Zeichenstrings gespeichert sind, und bei Übereinstimmung in das entsprechende Wörterbuch übertragen werden.

Zusammenfassung

Verfahren zur Bildung und/oder Aktualisierung von Wörterbüchern zum automatischen Adreßlesen

Es werden die vom OCR-Leser erzielten Leseergebnisse einer vereinbarten Anzahl von Sendungsbildern, unterteilt in eindeutig gelesene und zurückgewiesene Leseergebnisse zwischengespeichert.

Dann werden Klassen von Wörtern oder zusammengehörenden Wortgruppen der zwischengespeicherten und zurückgewiesenen Leseergebnisse, bestehend jeweils aus n Adreßwörtern, $n = 1,2,...,a$, mit den Wortabständen m, $m = 0,1,...,b$, gebildet, die bezogen auf jeweils einen bestimmten n- und m-Wert untereinander ein bestimmtes Ähnlichkeitsmaß nicht unterschreiten. Mindestens Repräsentanten derjenigen Klassen, deren Häufigkeit einen festgelegten Wert überschreiten, werden in das oder die Wörterbücher der zugeordneten Adreßbereiche aufgenommen.

Figur 1

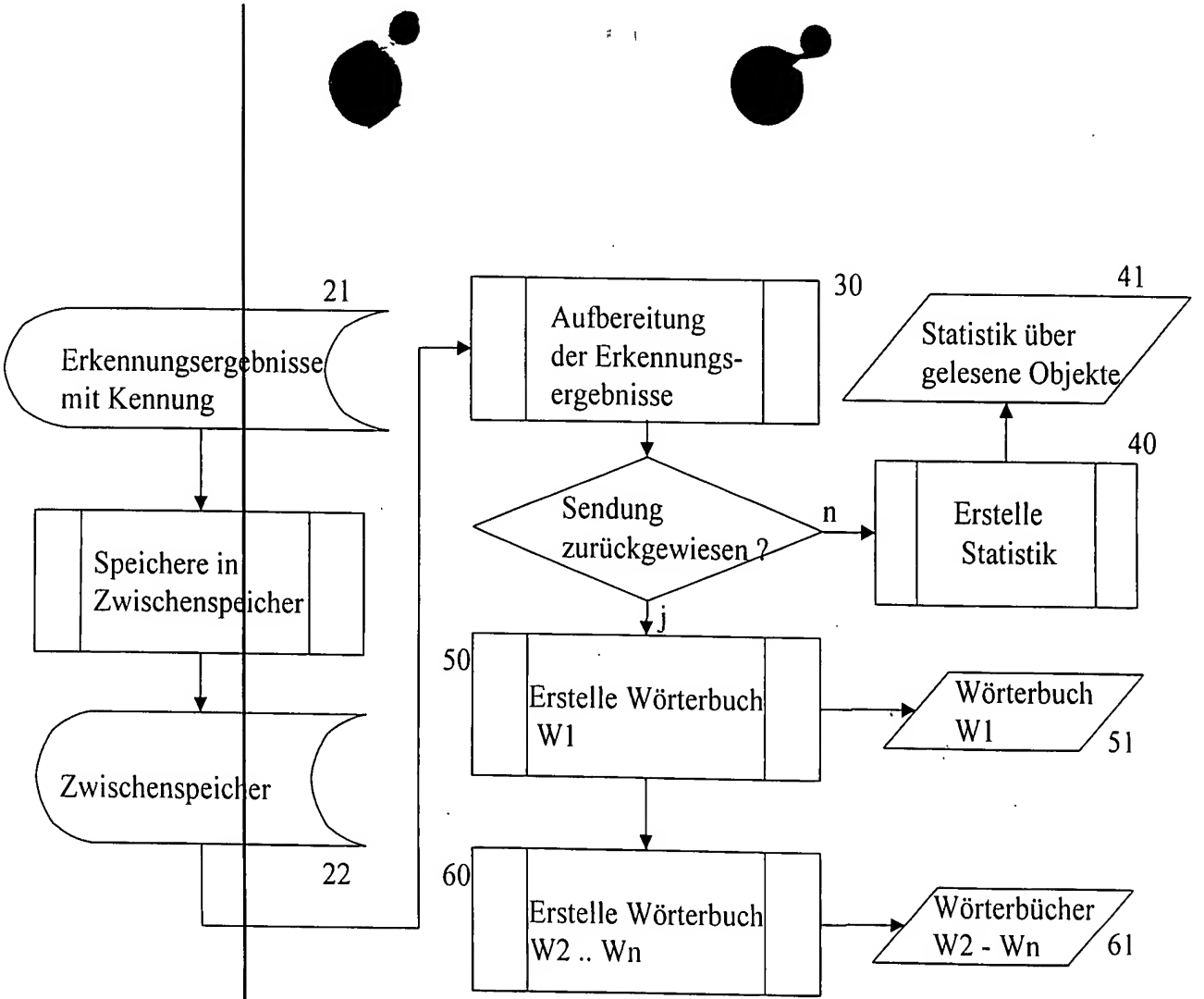

21
Erkennungsergebnisse mit Kennung
Speichere in Zwischenspeicher
Zwischenspeicher
22
30
Aufbereitung der Erkennungs-ergebnisse
Sendung zurückgewiesen ?
n
j
41
Statistik über gelesene Objekte
40
Erstelle Statistik
50
Erstelle Wörterbuch W1
Wörterbuch W1
51
60
Erstelle Wörterbuch W2 .. Wn
Wörterbücher W2 - Wn
61

FIG 1

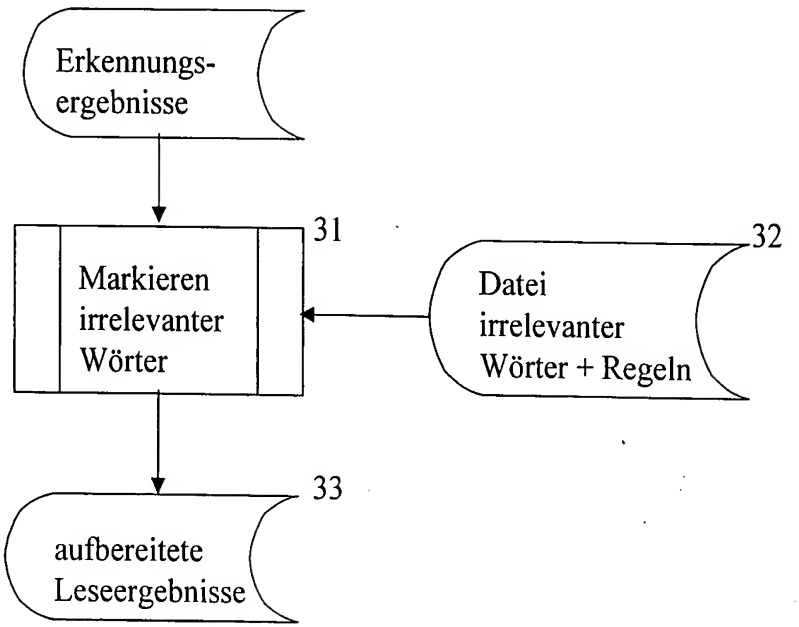
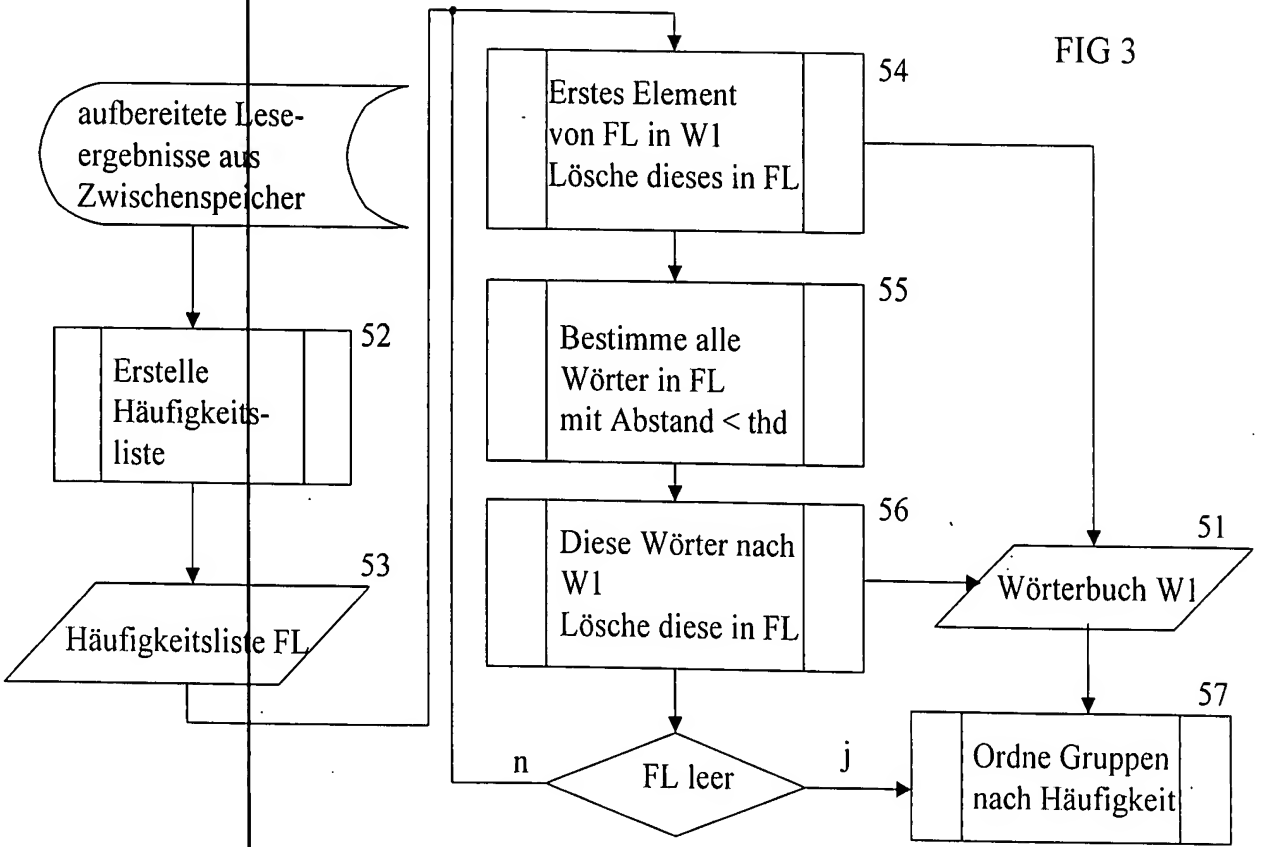

Erkennungs-
ergebnisse
31
Markieren
irrelevanter
Wörter
32
Datei
irrelevanter
Wörter + Regeln
33
aufbereitete
Leseergebnisse

FIG 2

FIG 3


aufbereitete Lese-ergebnisse aus Zwischenspeicher
52
Erstelle Häufigkeits-liste
53
Häufigkeitsliste FL
54
Erstes Element von FL in W1
Lösche dieses in FL
55
Bestimme alle Wörter in FL mit Abstand < thd
56
Diese Wörter nach W1
Lösche diese in FL
n
FL leer
j
51
Wörterbuch W1
57
Ordne Gruppen nach Häufigkeit

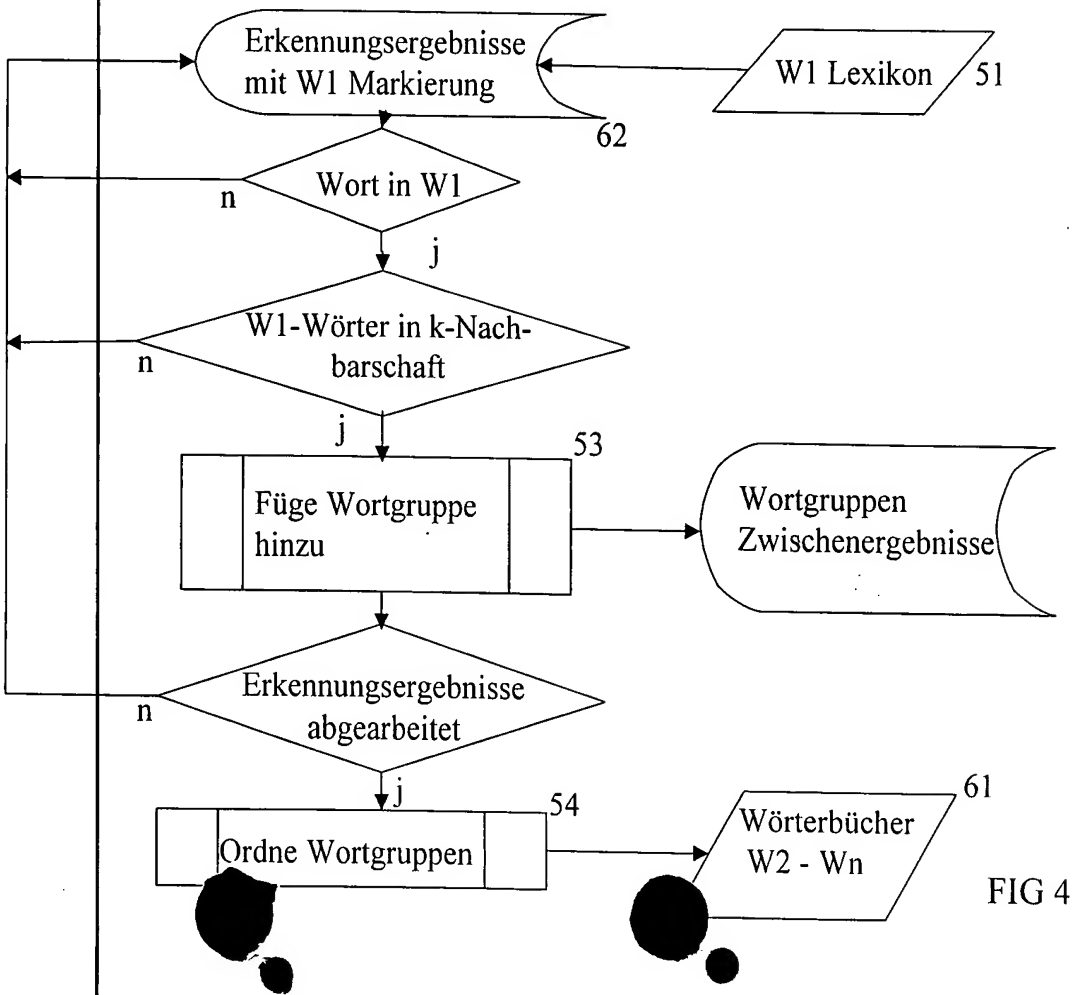

Erkennungsergebnisse mit W1 Markierung
62
W1 Lexikon
51
Wort in W1
n
j
W1-Wörter in k-Nachbarschaft
n
j
53
Füge Wortgruppe hinzu
Wortgruppen Zwischenergebnisse
Erkennungsergebnisse abgearbeitet
n
j
54
Ordne Wortgruppen
61
Wörterbücher W2 - Wn

FIG 4

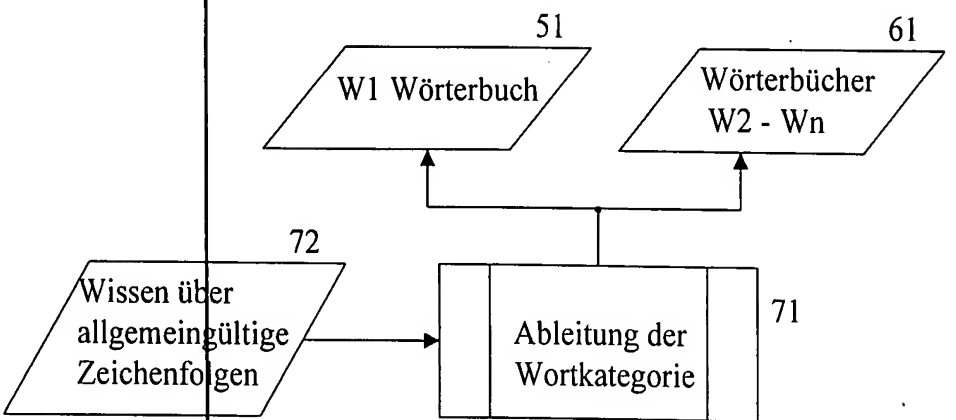

51
W1 Wörterbuch
61
Wörterbücher
W2 - Wn
72
Wissen über
allgemeingültige
Zeichenfolgen
Ableitung der
Wortkategorie
71

FIG 5